

On Things Not Seen

Marek Kimmel

Abstract Some statistical observations are frequently dismissed as “marginal” or even “oddities” but are far from such. On the contrary, they provide insights that lead to a better understanding of mechanisms which logically should exist but for which evidence is missing. We consider three case studies of probabilistic models in evolution, genetics and cancer. First, ascertainment bias in evolutionary genetics, arising when comparison between two or more species is based on genetic markers discovered in one of these species. Second, quasistationarity, i.e., probabilistic equilibria arising conditionally on non-absorption. Since evolution is also the history of extinctions (which are absorptions), this is a valid field of study. Third, inference concerning unobservable events in cancer, such as the appearance of the first malignant cell, or the first micrometastasis. The topic is vital for public health of aging societies. We try to adhere to mathematical rigor, but avoid professional jargon, with emphasis on the wider context.

1 Introduction

This essay attempts to persuade the Reader that statistical observations that may be dismissed as “marginal” or even “oddities” are far from such. On the contrary, they provide insights that lead to a better understanding of mechanisms which logically should exist but for which evidence is (and likely has to be) missing. To remain focused, we adhere to probabilistic models in evolution, genetics and cancer, disciplines in which the author claims expertise. The paper includes three case studies. First, ascertainment bias in evolutionary genetics, arising when comparison between two or more species is based on genetic markers discovered in one of these species.

M. Kimmel (✉)

Department of Statistics, Rice University, 6100 Main Street, Houston, TX 77005, USA
e-mail: kimmel@rice.edu

M. Kimmel

Systems Engineering Group, Silesian University of Technology,
Akademicka 16, 44-100 Gliwice, Poland

Second, quasistationarity, i.e., probabilistic equilibria arising conditionally on non-absorption. Since evolution is the history of extinctions (which are absorptions), this is a valid field of study. Third, inference concerning unobservable events in cancer, such as the appearance of the first malignant cell, or the first micrometastasis. The topic is vital for public health, particularly in aging societies. We try to adhere to mathematical rigor wherever needed and to provide references. Discussion concerns the wider context and philosophical implications.

2 Ascertainment Bias in Evolutionary Genetics

It has been observed that in evolutionary comparisons of Species 1 and 2, it is easy to err by using markers that were discovered in Species 1 and then sampled (“typed”) in Species 1 and 2. Genetic markers have to exhibit among-individual variation to be useful and therefore if a marker is discovered in Species 1, then on the average it is more variable in Species 1 than in Species 2. Variability of markers serves as a proxy for the rate of nucleotide substitution, which in turn may be a proxy for the rate of evolution. For this reason, if Species 1 and 2 descend from a common ancestral species, such as Human and Chimpanzee, and markers discovered in Species 1 (Human, for example) are employed, then we may deduce that Human has been evolving faster than its sister species Chimpanzee, when in fact it has not [2, 7, 23]. One remedy for this effect (being a form of the ascertainment bias) is to also use markers discovered in Species 2 and compare the outcomes in both cases. However, how to analyze such data and what inferences might be drawn? Li and Kimmel [19] demonstrate that this is quite complicated and that conclusions may be far from obvious.

2.1 *Microsatellite DNA and Divergence of Human and Chimpanzee*

Microsatellite loci are stretches of repeated DNA motifs of length of 2–6 nucleotides. An example is a triplet repeat (motif of length 3) with allele length $X = 4$ (motif repeated 4 times)

$$\dots |ACG|ACG|ACG|ACG| \dots .$$

Mutations in such loci usually have the form of expansions or contractions occurring at a high rate, $\nu \sim 10^{-3}$ – 10^{-4} per generation. More specifically,

$$X \longrightarrow X + U \tag{1}$$

where U is an integer-valued random variable, at times constituting a Poisson process with intensity ν . Mutations in this Stepwise Mutation Model (SMM), mathematically

form an unrestricted random walk (see e.g., [9]). Microsatellites are highly abundant in the genome. They are also highly polymorphic (variable). Applications of microsatellites include: forensics (identification), mapping (locating genes), and evolutionary studies.

A microsatellite locus can be considered to have a denumerable set of alleles indexed by integers. Two statistics can summarize the variability at a microsatellite locus in a sample of n chromosomes: The estimator of the genetic variance

$$\hat{V}/2 = \sum_{i=1}^n (X_i - \bar{X})^2 / (n - 1), \tag{2}$$

where $X_i = X_i(t)$ is the length of the allele in the i th chromosome in the sample and \bar{X} is the mean of the X_i

$$V(t) = E(\hat{V}) = E[(X_i - X_j)^2], \tag{3}$$

and X_i and X_j are exchangeable random variables representing the lengths of two alleles from the population [17]; and the estimator of homozygosity

$$\hat{P}_0 = (n \sum_{k=1}^K p_k^2 - 1) / (n - 1), \tag{4}$$

where p_k denotes the relative frequency of allele k in the sample

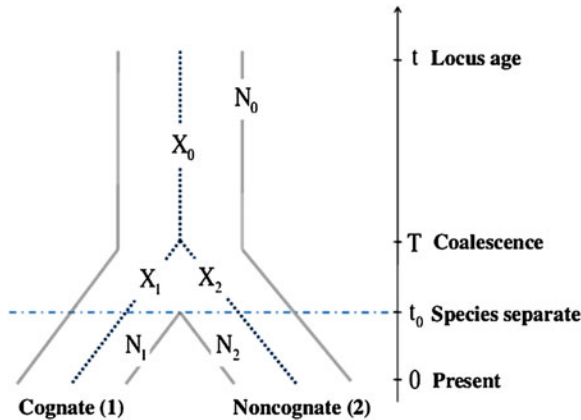


Fig. 1 Evolutionary history of a locus in two species. Demographic scenario employed in the mathematical model and simuPOP simulations. Notation: N_0 , N_1 , and N_2 , effective sizes of the ancestral, cognate, and noncognate populations, respectively; X_0 , X_1 , and X_2 , increments of allele sizes due to mutations in the ancestral allele, in chromosome 1 and in chromosome 2, respectively. From Ref. [19]

$$P_0(t) = E(\hat{P}_0) = \Pr[X_i(t) = X_j(t)]. \quad (5)$$

Random variables X_i are exchangeable but not independent.

Li and Kimmel [19] considered evolutionary history of a locus in two species. They employed the following demographic scenario in the mathematical model and simuPOP [20] simulations (Fig. 1). At time t before present (time is counted in reverse direction), a microsatellite locus is born in an ancestral species. At time t_0 , the ancestral species splits into species 1 (called cognate) and species 2 (called non-cognate). Notation: N_0 , N_1 , and N_2 , are effective lengths of the ancestral, cognate, and non-cognate populations, respectively; X_0 , X_1 , and X_2 are increments of allele lengths due to mutations in the ancestral allele, in chromosome 1 sampled at time 0 (present) from cognate population 1 and in chromosome 2, sampled from the non-cognate population 2.

2.2 Ascertainment Bias versus Drift and Mutation

In the random walk-like SMM model of mutation, a good measure of variability at a microsatellite locus is the length (repeat count) in a randomly sampled individual. Let us suppose that we discover a sequence of short motif repeats in the cognate species 1 and if its number of repeats Y_1 is greater or equal the threshold value x , we retain this microsatellite (we say we *discovered* it). Then we find a homologous microsatellite in species 2, i.e., microsatellite which is located in the same genomic region (technically, flanked by sequences of sufficient similarity), provided such microsatellite can be found. We take samples of microsatellite lengths from species 1 and 2, and consider their lengths to be realizations of random variables Y'_1 and Y_2 , respectively. We then consider the difference

$$D = E[Y'_1 | Y_1 \geq x] - E[Y_2 | Y_1 \geq x].$$

Other things being equal, D is a manifestation of the ascertainment bias and is likely to be positive. However, things may not be entirely equal. For example, if species 1 has a lower mutation rate than species 2, then its microsatellites will tend to have lower maximum length, which may reduce D . On the other hand, if, say, species 2 consistently has had a smaller population size, then genetic drift might have removed some of the variants and now species 2 microsatellites will have lower maximum length, which may inflate D . Li and Kimmel [19] carried out analytical and simulation studies of D under wide range of parameter values and obtained very good agreement of both techniques (Fig. 2). Briefly, as explained already, the observed difference D in allele lengths may be positive or negative depending on relative mutation rates and population sizes in the species 0 (ancestral), 1, and 2. In conclusion, mutation rate and demography may amplify or reverse the sampling (ascertainment) bias. Other effects were studied by different researchers. For example, Vowles and Amos [23] underscore the effects of upper bounds of repeat counts. An exhaustive discussion is found in Ref. [19].

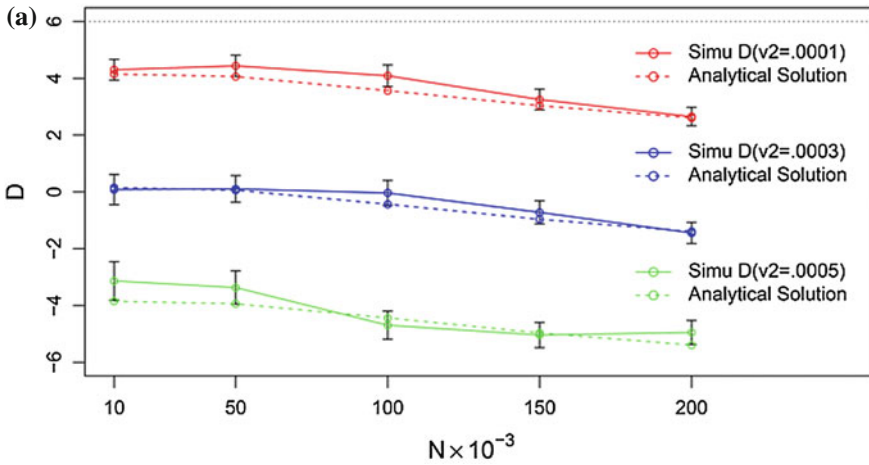


Fig. 2 Observed difference D in allele sizes may be positive or negative. Comparison of simuPOP simulations with computations based on Eq. (15). **a** Values of D for the basic parameter values $b_0 = b_1 = b_2 = b = 0.55$, $v_0 = v_1 = v = 0.0001$, $t_0 = 2 \times 10^5$ generations, and $t = 5 \times 10^5$ generations, with the effective sizes of all populations concurrently varying from 2×10^4 to 4×10^5 individuals and with mutation rates v_2 varying from v to $5v$. **b** Values of D for the basic parameter values $b_0 = b_1 = b_2 = b = 0.55$, $v_0 = v_2 = v = 0.0001$, $t_0 = 2 \times 10^5$ generations, and $t = 5 \times 10^5$ generations, with the effective sizes of all populations concurrently varying from 2×10^4 to 4×10^5 individuals and with mutation rates v_1 varying from v to $5v$ (assuming 20 years per generation). From Ref. [19]

2.3 Hominid Slowdown and Microsatellite Statistics

Li and Kimmel [19] considered evidence for and against the so-called hominid slowdown (as discussed e.g., in Bronham et al. 1996), the observation that as the great apes become closer to the Human lineage, their nucleotide substitution rates (rates of point mutations in the genome) decrease. Consistent with this, Human and Human ancestors are expected to have slower substitution rates than Chimpanzee and its ancestors (following the divergence from the common ancestral species about 7 million years ago). Is this also true of microsatellite loci? Different molecular mechanisms shape these two types of mutations. Nucleotide substitutions result from random errors in DNA replication, which then may not be repaired, but also may lead to dysfunctional proteins which will be eliminated from the population by natural selection (as discussed e.g., in [10]). Microsatellite mutations, as explained already, result from replicase slippage. Most microsatellites are located in noncoding regions and therefore are considered selectively neutral.

The study [19] involves a reconstruction of the past demography of Human and its ancestors as well as hypothetical demography of Chimpanzee and its ancestors, including migrations of Human from its ancestral African territory and resulting population growth interrupted by recent glaciations and other events. Without getting

into technical details, the conclusion is that microsatellite mutation rate is likely to be higher in Human than in Chimpanzee. It is interesting to observe that also the regulatory sites in the genome usually have the form of simple repeats (albeit interrupted) and vary quite considerably among species of mammals (as reviewed e.g., in Ref. [13]). It is possible to further hypothesize that evolution in higher mammals chose the path of regulation of gene expression as opposed to modification of the amino acid sequences in proteins; possible reason being that these latter might be too slow.

3 Quasistationarity in Genome Evolution

Let us consider an effect which is important if extinctions are indeed common in evolution. Suppose that a proliferating population has a random component of such nature that it leads any lineage to extinction with probability 1. On the other hand, proliferation is sufficiently fast to make up for extinction so that the non-extinct part may persist indefinitely. The long-term distribution of types of individuals in the population conditional on non-extinction, if such distribution exists, is called the quasistationary distribution. Quasistationarity in a more general sense has been studied by mathematicians for a long time; relevant literature has been collected by Pollet [21]. Here we will limit ourselves to an example from cell biology concerning gene amplification, based on an experiment pioneered by Schimke [22], with mathematical model developed by Kimmel and Axelrod [15] and then generalized by Kimmel [14] and Bansaye [3]. Let us notice that extinction causes information about evolution of the population to be scrambled. Therefore, if quasistationary distributions are interpreted as if they were ordinary stationary distributions, the conclusions may be paradoxical or misleading.

3.1 *Gene Amplification in Cancer and Schimke's Experiments*

One of the prevalent types of rearrangements in human cancer genome is gene amplification, i.e., increase of the number of gene copies in cells beyond the usual diploid complement. Some examples have been provided by [1], but the phenomenon is quite common, usually appearing under the guise of copy number variation (CNV; Fig. 3).

Classical experiments demonstrating gene amplification and its connection with drug resistance have been carried out in Schimke [22]. The gist of the experimental data can be described as follows. After passaging surviving cultured cells to ever increasing levels of methotrexate (MTX) over the period of the order of 10 Msec = 5 month, it was possible to evolve cells that were resistant to extremely high doses of MTX (Fig. 4). When the cells were put back into no MTX medium, they were observed to lose resistance within about 100 cell doublings (some cultures did not, but we sweep these under the rug for now).

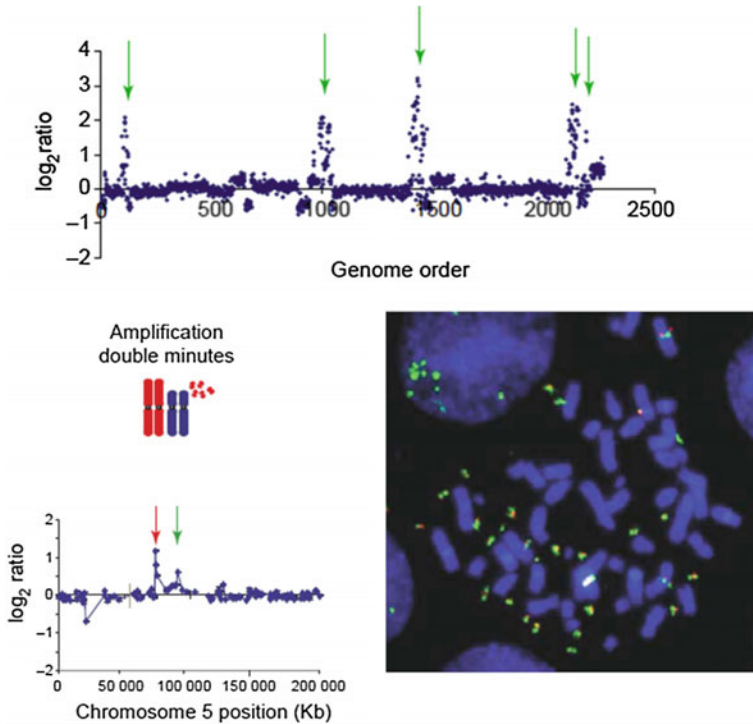
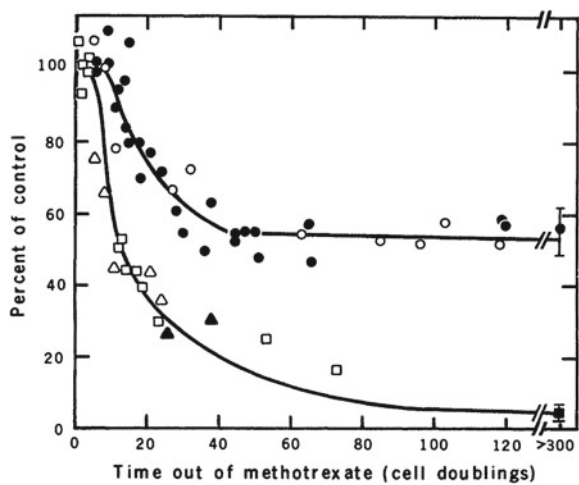


Fig. 3 Cytogenetics of gene amplification. Amplified DNA can be present in various forms including double minutes. A two-chromosome genome is depicted (*top of the figure*). Examples of array CGH copy number profiles (*bottom left*; plotted as the normalized log₂ ratio) are shown with corresponding FISH pictures (*bottom right*) of the cells using BAC clones from the region of the amplicon indicated by the *red and green arrows*. Many red and green signals can be seen in the double minutes in a methotrexate-resistant human cell line. **From Ref. [1]**

Fig. 4 Loss of resistance in Schimke's experiments. Cells resistant to MTX are exposed to nonselective conditions. Some cell lines lose resistance completely (*circles*), while other only partially (*squares and triangles*). **From Ref. [5]**



Schimke discovered, using techniques available at that time, that the highly resistant cell had, besides the usual chromosomes, small extrachromosomal DNA elements roughly dicentric (he named them the “double minute chromosomes” or DM for short) that contained extra copies of the dihydrofolate reductase (DHFR) gene, that confers resistance to MTX [1]. It became clear that the increased resistance was due to amplification of the DHFR gene. But how did the amplified copies get there? Clearly a supercritical process of gene copy proliferation was at play. However, how did the cells know to multiply gene copies? The ghost of Lamarck knocked at the door.

3.2 *Probabilists to Rescue*

Fortunately for the common sense, Kimmel and Axelrod [15] conceived an idea consistent with the neo-Darwinian paradigm (despite appearances, this sentence is not necessarily an oxymoron). The hypothesis can be stated as follows:

- Increased resistance is correlated with increased numbers of gene copies on double minute chromosomes (DM).
- The number of DHFR genes on double minutes in a cell may increase or decrease at each cell division. This is because double minutes do not have centromeres, which are required to faithfully segregate chromosomes into progeny cells.
- The process of DM proliferation in cells is subcritical, since the DM do not efficiently replicate. Therefore cells grown in the absence of the drug gradually lose resistance to the drug, by losing extra gene copies.

The following model has been constructed by Kimmel and Axelrod [15].

- Galton-Watson process of gene amplification and deamplification in a randomly chosen line of descent (Fig. 5).
 - Double minute chromosomes replicate irregularly
 - Upon cell division, DMs are asymmetrically assigned to progeny cells.
- The process is subcritical, i.e., the average number of DMs at division is less than twice that number assigned to the cell at birth. This is consistent with imperfect replication and segregation of DMs.

Hypotheses of the model explain why, under nonselective conditions, the number of DMs per cell decreases which causes gradual loss of resistance (Fig. 5). In other words, zero is an absorbing state for the number of DMs. However, under selective conditions, only the cells with nonzero DM count survive. Therefore, conditionally on nonabsorption (non-extinction of the DMs), according to the Yaglom theorem for subcritical branching processes, the number of DMs per cell converges in distribution to a quasistationary distribution.

Specifically, suppose that proliferation of DMs from one cell generation to another, in a randomly selected ancestry line of cells is described by a Galton-Watson branch-

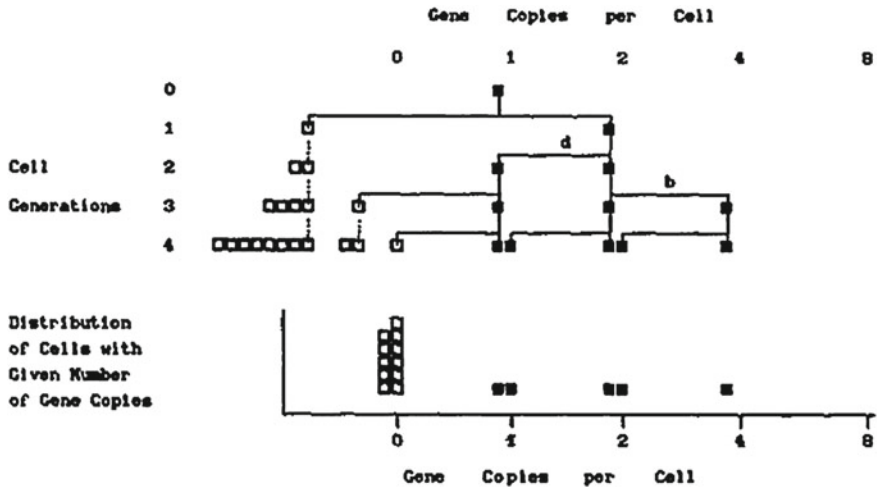


Fig. 5 A simplified view of gene amplification and deamplification process. Each cell with at least one gene copy can give rise to 2 progeny cells, each of which with probability b has amplified (doubled) count of DM gene copies, with probability d has deamplified (halved) count, or with probability $1 - b - d$, the same number. Halving of a single DM results in 0 DMs. Histogram at the bottom shows the resulting distribution of gene copies per cell in the fourth generation. **From** Ref. [15]

ing process with the number of “progeny” of a DM is a nonnegative integer random variable with generic probability generating function (pgf) $f(s)$, under the usual conditional independence hypotheses. As already noticed, this process is subcritical, i.e., $m = f'(1-) < 1$. Let Z_n denote the number of DMs in generation n and let $f_n(s)$ denote the pgf of Z_n .

Yaglom Theorem (see e.g., Theorem 4 in Kimmel and Axelrod [16]) *If $m < 1$, then $P[Z_n = j | Z_n > 0]$ converges, as $n \rightarrow \infty$ to a probability function whose pgf $\mathcal{B}(s)$ satisfies the equation*

$$\mathcal{B}[f(s)] = m\mathcal{B}(s) + (1 - m).$$

Also,

$$1 - f_n(0) \sim \frac{m^n}{\mathcal{B}'(1-)}, \quad n \rightarrow \infty.$$

Yaglom limit is also an example of a *quasistationary* distribution, say $\mu(x)$, which in a general Markov chain can be defined via the following condition

$$\mu(x) = \frac{\sum_{y \geq 1} \mu(y) P_y[X(t) = x]}{\sum_{y \geq 1} \mu(y) P_y[X(t) \neq 0]},$$

where $P_y[X(t) = x]$ is the transition probability matrix.

Let us suppose now that cell population has been transferred to MTX-free medium at generation $n = N$. Based on the Yaglom Theorem, the fraction of resistant cells decreases roughly geometrically

$$1 - f_n(0) \sim \frac{m^{n-N}}{B'(1-)}, \quad n > N,$$

while $\{Z_n | Z_n > 0\}$ remains unchanged. Moreover, if $2m > 1$, then the net growth of the resistant population is observed also at the selection phase ($n \leq N$).

Loss of DMs in non-selective conditions has been visualized experimentally [5]. Population distribution of numbers of copies per cell can be estimated by flow cytometry. Proportion of cells with amplified genes decreases with time (Fig. 6). Shape of the distribution of gene copy number in the subpopulation of cells with amplified genes appears unchanged as resistance is gradually lost.

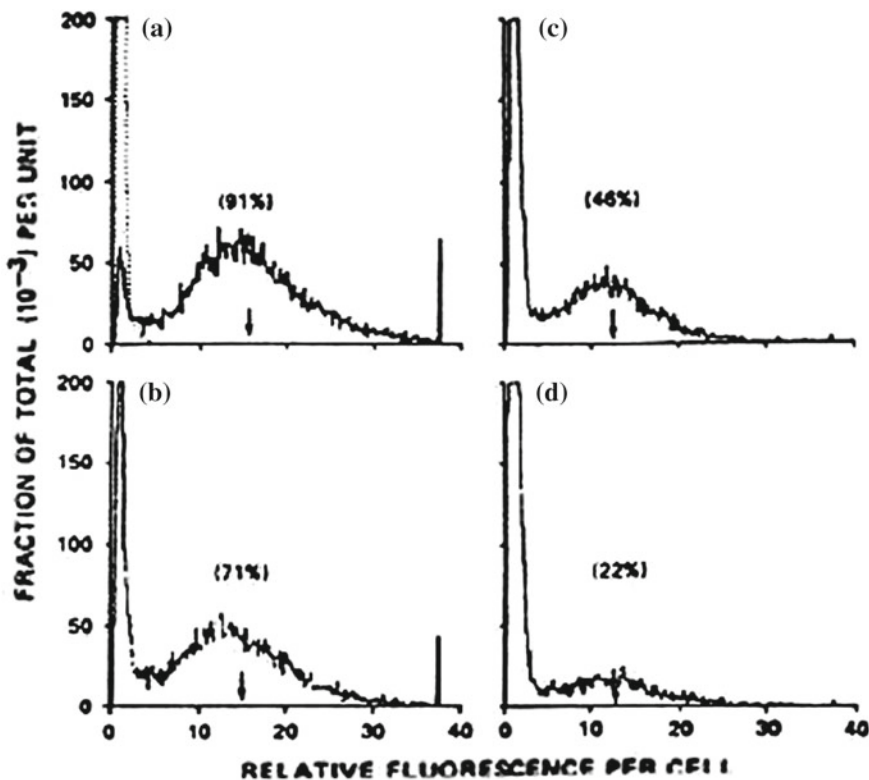


Fig. 6 Loss of resistance visualized by flow cytometry. Population distribution of numbers of copies per cell can be estimated by flow cytometry. Proportion of cells with amplified genes decreases with time. Shape of the distribution of gene copy number in the subpopulation of cells with amplified genes appears unchanged as resistance is gradually lost. **From Ref. [5]**

An important finding is that if $2m > 1$, i.e., if absorption is not too fast, then cell proliferation outweighs the loss of cell caused by the selective agent (MTX) and the resistant subpopulation grows in absolute numbers also under selective conditions (when $n \leq N$; details in the original paper and the book).

A more general mathematical model of replication of “small particles” within “large particles” and of their asymmetric division (“Branching within branching”) has been developed by Kimmel [14] and followed up by Bansaye [3]. It is interesting to notice that quasistationary distributions are likely to generate much heterogeneity. An example is provided by large fluctuations of the critical Galton-Watson process before extinction; see Wu and Kimmel [24].

3.3 Quasistationarity and Molecular Evolution

An observation can be made that trends observed in molecular evolution can be misleading, if they are taken at their face value and without an attempt to understand their underlying “mechanistic” structure. It may be concluded, looking at the evolution of resistance in cells exposed to MTX that there exists something in the MTX that literally leads to an increase of the number of DM copies. So, gene amplification is “induced” by MTX. Only after it is logically deduced that DMs have to undergo replication and segregation and assuming that both these processes are less orderly in DMs than in the “normal” large chromosomes, the conclusion concerning the true nature of the process (selection superimposed on *subcritical* branching) follows by the laws of population genetics.

4 Unobservables in Cancer

Early detection of cancer by mass screening of at risk individuals remains one of the most contentious issues in public health. We will mainly use lung cancer (LC) as an example. The idea is to identify the “at risk” population (smokers in the LC case), and then to apply an “early detection” procedure (CT-scan in the LC case), periodically, among the members of the “at risk” population. By treating the early (and implicitly, curable) cases discovered this way, a much higher cure rate is assured than that of spontaneously appearing symptomatic cases. Is this reasoning correct? Two types of arguments have been used to question the philosophy just described. On one hand, part of the early detection may constitute overdiagnosis. Briefly, by the effect known from the renewal theory, a detection device with less than perfect sensitivity, placed at a fixed point in time and confronted with examined cases “flowing in time”, preferentially detects cases of longer duration, i.e. those for which the asymptomatic early disease is more protracted. This effect is known as the length-biased sampling (discussion in Ref. [12]). Its extreme form, called overdiagnosis, causes detection of cases that are so slow that they might show only at autopsy, or cases which look

like cancer but do not progress at all. Overdiagnosis, if it were frequent, would invalidate early detection: a large number of “early non-cancers” would be found and unnecessarily treated, causing increased morbidity and perisurgical mortality, without much or any reduction in LC death count.

On the other hand, the following scenario is possible, which also may invalidate screening for early detection, although for an opposite reason. If it happens that LC produces micrometastases, which are present when the primary tumor is of submillimeter size, then detection of even 2–3 mm tumors (achievable using CT) is futile, since the micrometastases progress and kill the patient whether the primary tumor has been detected or not.

How to determine if screening reduces number of LC deaths? The orthodox biostatistics approach is “empirical”. It consists of designing a two-arm RCT (screened versus non-screened high risk individuals) and comparing numbers of LC deaths in the two arms. This methodology is statistically sound, but it may be considered unethical. Patients in the control arm are denied potentially life-saving procedures. Those in the screened arm do not necessarily take advantage of the current state-of-art technology. Two sources of reduced contrast are: noncompliance in the screened arm and/or “voluntary” screening in the control arm. It has been claimed that the results of the Mayo Lung Project (MLP) 1967–1981 trial, which influenced recommendations not to screen for LC by chest X ray were simply due to lack of power to demonstrate mortality reduction by 5–10% which might be achievable using X ray screening [12]. Finally, the National Lung Screening Trial (NLST) in the USA, in which around 50,000 smokers took part, demonstrated that a series of three annual screenings followed by treatment of detected cases reduced mortality by about 20%. It has to be noted, that predictions of similar magnitude reduction obtained using modeling [18] have been almost universally disregarded by the medical community.

The NLST has left as many questions unanswered as it did answer. One of them is the choice of the “best” high-risk group for LC screening. Given limited resources, how to allocate them to subgroups of screenees so that the efficacy of a mass screening program is maximized. Even if the meaning of the term “efficacy” is clarified, it is still unknown who should be screened. Are these the heaviest smokers, the smokers who have smoked for the longest period of time, individuals with family history of lung cancer, or those with impaired DNA-repair capacity [11]? At what age does it make sense to start screening and how often should the individuals be screened? Answers to these questions require knowledge of the natural course of disease, which is exactly what is not observable (Fig. 7).

Arguably, modeling can help. If a model of carcinogenesis, tumor growth and progression (i.e., nodal and distant metastases) is constructed and validated and models of early detection and post-detection follow-up are added to it, then various scenarios of screening can be tested *in silico*. Another use of modeling is less utilitarian, but equally important. It can be called the inverse problem: How much is it possible to say about the course of cancer based on snapshots including the disease characteristics at detection? To what extent is the size of the primary tumor predictive of the progression of the disease? In [6] some inferences of this type have been made (Fig. 8).

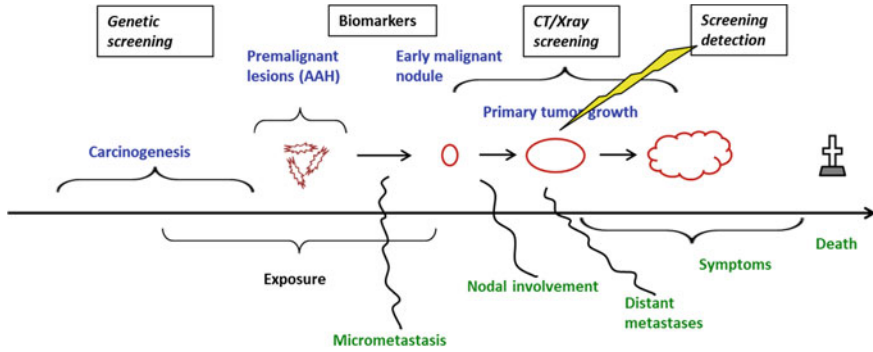


Fig. 7 Time lines of cancer progression and detection

Observed stage	True stage	Groups of predicted lung cancer tumor size						
		Total	TS*≤0.5	0.5<TS≤ 1	1<TS≤ 1.5	1.5<TS≤ 2	2<TS≤ 3	TS >3
N0M0		N=380017	N=3040	N=18301	N=47534	N=59685	N=100031	N=151426
	N0M0, n (%)	245871 (64.7)	2860 (94.1)	12151 (66.4)	20749 (43.7)	21735 (36.4)	53346 (53.3)	135030 (89.1)
	N1M0, n (%)	35342 (9.3)	122 (4.0)	3704 (20.2)	12000 (25.2)	9583 (16.1)	4084 (4.1)	5849 (3.9)
	M1, n (%)	98804 (26.0)	58 (1.9)	2446 (13.4)	14785 (31.1)	28367 (47.5)	42601 (42.6)	10547 (7.0)
N1M0**		N=321221	N=813	N=5500	N=16827	N=24513	N=62866	N=210702
	N1M0, n (%)	138412 (43.1)	150 (18.5)	506 (9.2)	3512 (20.9)	4769 (19.2)	12084 (19.2)	117391 (55.7)
	N1M1, n (%)	182809 (56.9)	663 (81.5)	4994 (90.8)	13315 (79.1)	19744 (80.8)	50782 (80.8)	93311 (44.3)
M1		N=732786	N=2058	N=13450	N=33883	N=54203	N=138575	N=490617
	M1, n (%)	732786 (100)	2058 (100)	13450 (100)	33883 (100)	54203 (100)	138575 (100)	490617 (100)

Fig. 8 Distributions of occult nodal and distant metastases in the simulated lung cancer patients (1988–1999) with stage N0M0, N1M0 and M1 stratified by tumor size. *TS, Primary tumor size (cm) in diameter **In SEER data, 7208 were N0M1, which is 9.7% of 74109 that had N and M staged. This stage is not modeled. From Ref. [6]

Figure 8 depicts distributions of undetected (so-called occult) nodal and distant metastases in the simulated lung cancer patients, fitting demographic and smoking patterns of the SEER database 1988–1999, detected with stage N0M0, N1M0 and M1, stratified by primary tumor size. N0 and N1 correspond to the absence and presence of lymph node metastasis, and M0 and M1 to the absence and presence of distant metastasis, respectively. In other words, modeling allows to estimate how many of lung cancers detected as belonging to a given category, in reality belong to different, prognostically less favorable, categories. The results show some unexpected trends. The most important are the three top rows of Fig. 8, which concern tumors detected without nodal or distant metastases (N0M0). These tumors, on the face of things,

offer best prognosis. Model predictions confirm this intuition, up to a point. Indeed up to the primary tumor size of about 1 cm, more than 50% of apparent NOM0 tumors are indeed NOM0. If they are detected at larger sizes, then only a minority are truly NOM0, and the rest have occult metastases. So, if these tumors below 1 cm are removed, there is a good chance the patient is cured. But, surprisingly, there is another turning point. At sizes above 2.5–3 cm, again majority of tumors are NOM0. Similar, though not as distinctive trend is present when we consider tumors detected as N1M0. Therefore, if a very large tumor is discovered without apparent nodal and distant metastasis and it is resectable, then the suggestion is that it might be resected for cure.

The explanation for this peculiar pattern is that if the rates of growth and progression (metastasizing) of tumors are distributed, then detection is “cutting out windows” in the distributions, through which the tumor population is observed. In the large primary tumor size category with no metastases observed, we deal with the fast growing, slowly metastasizing subset. Other subpopulations simply present with metastasis when the primary tumor is large, become symptomatic and quickly progress to death. So, active detection leads to biased TNM distributions, with the bias sometimes being non-intuitive.

Mathematical models of the NLST trial predicted its outcome in two publications, one in 2004 [18] and the other in 2011 ([8]; submitted for publication before the NLST outcome was announced), using two different modeling approaches. As stated already, at that time these papers were universally ignored.

5 Discussion

What is the role and use of statistics as a profession (science?) and of statisticians as professionals (scientists?). In minds of other scientists (physicists, biologists or physicians) statistics is mainly perhaps a useful, but strictly confirmatory field. What is expected of a collaborating statistician is the “ p -value” or the sample size needed to obtain a given “power” of a test as required by the funding agencies. However, one may reflect on how many useful and deep scientific concepts and techniques are statistical in nature. Some of them have been revolutionary. We may list some with biological applications: Fluctuation Analysis (FA) in cell biology, Moolgavkar-Knudson (M-K) model of carcinogenesis, Wright-Fisher (W-F) model in population genetics, Capture-Recapture (C-R) method in ecology, Maximum Likelihood (ML) and Least Squares (LS) methods in molecular phylogenetics, and other. However, let us notice that these methods are based on models that include structural features of the biological nature of the phenomenon in question. Some of these are unobservable, such as mutations in cells in FA, stage transition times in M-K, segregation of chromosomes to progeny in W-F, collections of individuals in C-R and ancestral nodes in phylogenetics.

Arguably, statistics is most useful, when it considers phenomena in a “gray zone” such as inference on the unseens, i.e., processes that we believe are real, but which cannot be directly observed. Three phases of scientific inquiry, are usually present:

1. Initially, when there is little or no data; the unseens are not suspected to exist,
2. Existence of the unseens is revealed through progress in data collection and logical analysis,
3. Further progress may lead to resolution of the unseen by a reductionist approach.

Examples considered in the essay involve analyses in Phase 2. Each involves unseens that may become observable at some time. Also, each required construction of a new model based on inferred biology of the process. In addition, each of the examples includes a statistical sampling mechanism, which introduces a bias (we may call it the ascertainment bias). The role of the model is among other, to understand and counter the bias. Arguably, this is the true purpose of statistical analysis.

References

1. Albertson DG (2006) Gene amplification in cancer. *Trends Genet* 22:447–455
2. Amos W et al (2003) Directional evolution of size coupled with ascertainment bias for variation in drosophila microsatellites. *Mol Biol Evol* 20:660–662
3. Bercu B, Blandin V (2014) Limit theorems for bifurcating integer-valued autoregressive processes. *Statistical inference for stochastic processes*, pp 1–35
4. Bromham L, Rambaut A, Harvey PH (1996) Determinants of rate variation in mammalian DNA sequence evolution. *J Mol Evol* 43:610–621
5. Brown PC, Beverley SM, Schimke RT (1981) Relationship of amplified dihydrofolate reductase genes to double minute chromosomes in unstably resistant mouse fibroblast cell lines. *Mol Cell Biol* 1:1077–1083
6. Chen X et al (2014) Modeling the natural history and detection of lung cancer based on smoking behavior. *PloS one* 9(4):e93430
7. Cooper G, Rubinsztein DC, Amos W (1998) Ascertainment bias cannot entirely account for human microsatellites being longer than their chimpanzee homologues. *Hum Mol Genet* 7:1425–1429
8. Foy M et al (2011) Modeling the mortality reduction due to computed tomography screening for lung cancer. *Cancer* 117(12):2703–2708
9. Goldstein DB, Schlotterer C (1999) Microsatellites: evolution and applications, pp 1–368
10. Gorlov IP, Kimmel M, Amos CI (2006) Strength of the purifying selection against different categories of the point mutations in the coding regions of the human genome. *Hum Mol Genet* 15:1143–1150
11. Gorlova OY et al (2003) Genetic susceptibility for lung cancer: interactions with gender and smoking history and impact on early detection policies. *Hum Hered* 56:139–145
12. Gorlova OY, Kimmel M, Henschke C (2001) Modeling of long-term screening for lung carcinoma. *Cancer* 92:1531–1540
13. Iwanaszko M, Brasier AR, Kimmel M (2012) The dependence of expression of NF- κ B-dependent genes: statistics and evolutionary conservation of control sequences in the promoter and in the 3' UTR. *BMC Genomics* 13:182
14. Kimmel M (1997) Quasistationarity in a branching model of division-within-division. In: *Classical and modern branching processes* (Minneapolis, MN, 1994), pp 157–164. Springer, New York. IMA Vol Math Appl 84

15. Kimmel M, Axelrod DE (1990) Mathematical models of gene amplification with applications to cellular drug resistance and tumorigenicity. *Genetics* 125:633–644
16. Kimmel M, Axelrod DE (2015) *Branching processes in biology* (2nd edn, extended). Springer, Heidelberg
17. Kimmel M et al (1996) Dynamics of repeat polymorphisms under a forward-backward mutation model: within-and between-population variability at microsatellite loci. *Genetics* 143:549–555
18. Kimmel M, Gorlova OY, Henschke CI (2004) Modeling lung cancer screening. *Recent advances in quantitative methods in cancer and human health risk assessment*, pp 161–175
19. Li B, Kimmel M (2013) Factors influencing ascertainment bias of microsatellite allele sizes: impact on estimates of mutation rates. *Genetics* 195:563–572
20. Peng B, Kimmel M (2005) simuPOP: a forward-time population genetics simulation environment. *Bioinformatics* 21:3686–3687
21. Pollett PK (2014) Quasi-stationary distributions: a bibliography. <http://www.maths.uq.edu.au/~pkp/papers/qsds/qsds.pdf>
22. Schimke RT (ed) (1982) *Gene amplification*, vol 317. Cold Spring Harbor Laboratory, New York
23. Vowles EJ, Amos W (2006) Quantifying ascertainment bias and species-specific length differences in human and chimpanzee microsatellites using genome sequences. *Mol Biol Evol* 23:598–607
24. Wu X, Kimmel M (2010) A note on the path to extinction of critical Markov branching processes. *Statist Probab Lett* 80:263–269